Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 154 (2016) 1267 – 1274

Procedia Engineering

www.elsevier.com/locate/procedia

12th International Conference on Hydroinformatics, HIC 2016

Understanding water level residuals in Malacca Strait using genetic programming

Serene Hui Xin Tay[a,]*, Vladan Babovic[a]

[a]*Department of Civil and Environmental Engineering, National University of Singapore, E1-08-25, 1 Engineering Drive 2, Singapore 117576*

Abstract

Hydrodynamics are highly complex in Malacca Strait as it is where tides from Indian Ocean and South China Sea interact. Highly varying topography and geometry, river discharges from land and seasonal monsoon climate contribute further complication to the local flow dynamics and usually requires numerical model to resolve. However, no matter how well the numerical model is calibrated, residual will exist due to imperfect description of underlying physics and lack of high quality input data. Numerous studies have applied data-driven methods to correct numerical model prediction by forecasting the residuals, and shown that these methods are undeniably effective and efficient and being great value to more traditional modelling approaches. However, in complex hydrodynamic system of Malacca Strait, instead of simply treating numerical model residual as a numerical mismatch and addressing it as a time series problem by local correction, in this paper a more interesting and meaningful effort to uncover the underlying dynamics is attempted. This paper explores the ability of genetic programming to unearth the embedded components or dependencies of the numerical model residual in Malacca Strait.

Peer-review under responsibility of the organizing committee of HIC 2016

Keywords: genetic programming; water level; residuals; Malacca Strait

* Corresponding author. Tel.: +6565163650
E-mail address: serenetay@nus.edu.sg

Peer-review under responsibility of the organizing committee of HIC 2016
doi:10.1016/j.proeng.2016.07.458

1. Introduction

Geographically located between Andaman Sea and South China Sea, water levels in Malacca Strait are indirectly driven by tide and hydrodynamic components from the two oceans: Indian Ocean and Pacific Ocean (Fig 1). Numerical models such as Tay et al. [1], Tay et al. [2] and Tay et al. [3] have shown to describe tidal and non-tidal flows well in the strait. The so-called 'South China Sea Model Curvilinear' (SCSMC) [1] which is built in Delft3D modelling environment shows good tidal representation in South China Sea and Malacca Strait. Tay et al. [3] apply the same 2D barotropic model to simulate hydrodynamics in Malacca Strait that include both tidal and non-tidal flows by imposing additional meteorological forcing such as wind and pressure on the water surface. Significant improvement in non-tidal water level, also known as sea level anomalies (SLA) representation in the strait is demonstrated by inclusion of additional SLA forcing. This SLA forcing is prescribed in the form of water level at the model open boundary in Andaman Sea which is adjacent to Malacca Strait (Fig 1). Tay et al. [3] has described this additional SLA forcing as 'tilt', and could be derived from either a spatially larger model that covers the entire Indian Ocean or satellite altimetry data. The latter has shown to be giving better accuracy in SLA representation in Malacca Strait and details can be found in Tay et al. [2] and Tay et al. [3].

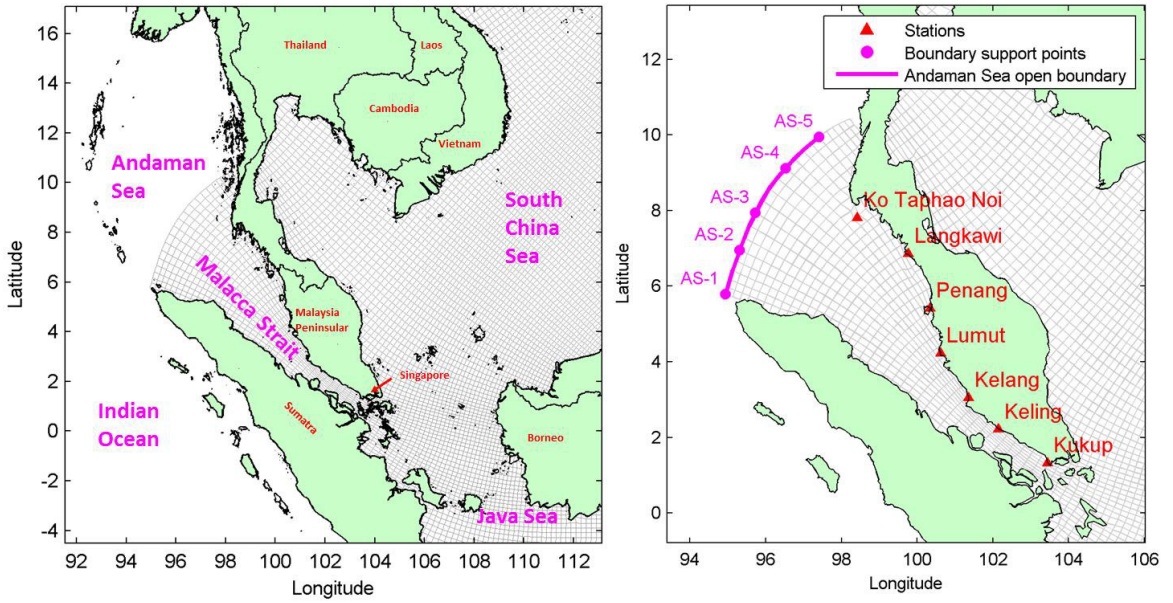

Fig 1. (a) Map of the Malacca Strait and its surrounding seas with SCSMC grid (grey); (b) location of Andaman Sea open boundary support points and UHSLC observation station

Nomenclature

s	station
t	time
$WL_{residual}$	residual water level
$WL_{obervation}$	observed water level
WL_{model}	model water level
AS	DUACS SLA
j	index of Andaman Sea open boundary support point
α	GP derived parameter
λ	GP derived parameter
γ	GP derived parameter

Although representation of hydrodynamic processes can be modelled well in Malacca Strait based on approaches presented in earlier studies described above, model residual of total water level representation remains significant in Malacca Strait. This will be shown in subsequent section of this paper (Table 2). Residual could be caused by baroclinic local effects such as river discharges which are not taken into account in the model. Other sources of inaccuracies are approximation error in bathymetry, grid resolution and boundary forcing. These result in residual being highly non-linear and complex. There have been studies focusing on residual prediction in order to correct and improve the water level prediction [4-6]. These data-driven approaches mainly rely on embedded temporal characteristic of historical time series of a component to be predicted, such as water level or SLA at a location of interest, to predict behaviour of the same component in the future. Based on remarkable results presented in these studies, data-driven methods are undeniably effective and efficient and being great value to more traditional modelling approaches. However, in complex hydrodynamic system of Malacca Strait, instead of simply treating numerical model residual as a numerical mismatch and addressing it as a time series problem by local correction, in this paper a more interesting and meaningful effort to uncover the underlying dynamics is attempted. One method that has been proven to be useful for determining the underlying dynamics is evolutionary computing based genetic programming (GP) [7-9] by describing dependent components of residuals explicitly. With a GP induced model that describes residuals well, correction can be made to the model water level for better prediction. Therefore the objective of this paper is to explore ability of GP to unearth the embedded components or dependencies of the numerical model residual in Malacca Strait, and improve water level prediction using residual correction model induced by GP.

2. Methodology

2.1. Genetic Programming

Genetic programming (GP) is an evolutionary algorithm technique introduced by Koza [7] to automatically solve problems without requiring the form or structure of the solution to be prescribed in advance. Differing from other evolutionary algorithm techniques such as genetic algorithm [10, 11] and evolution strategies [12, 13] which are typically applied to optimization problems, GP is suited for machine learning problems [14, 15]. Koza [7] and Keijzer and Babovic [16] described many seemingly different problems of various fields can be reformulated as problems of model (or program) induction. According to this GP paradigm is able of searching the space of possible models for an individual model that is highly fit for solving a particular problem. Babovic and Keijzer [8] further refine model induction application of GP as a knowledge discovery tool and develop physically sound empirical relationship using a dimensionally aware GP with potentially dependent input variables. One popular application of GP is referred to as symbolic regression [8, 17, 18]. Unlike linear or non-linear regression in which numeric coefficients of a predefined function are to be determined, GP is applied akin to symbolic regression to establish both the functional form and determine numeric coefficients simultaneously. GPKERNEL, a GP modeling tool developed by Babovic and Keijzer [8] is used as GP implementation tool in this paper. Details on the general preparatory steps and execution of GP can be found in Banzhaf et al. [15] and Keijzer and Babovic [16].

2.2. Data

Bulk of insitu observations used in this paper consist of 14 years (1993 – 2006) of daily water level data time series produced by University of Hawaii Sea Level Center (UHSLC) (http://uhslc.soest.hawaii.edu/) at seven stations: Ko Taphao Noi, Langkawi, Penang, Lumut, Kelang, Keling and Kukup in Malacca Strait (Fig 1).

As mentioned earlier, the SLA representation in Malacca Strait is significantly improved by applying tilt on the Andaman Sea open boundary in addition to tide and meteorological forcing in the model [2, 3]. Tilt is applied on five boundary support points i.e. AS-1, AS-2, AS-3, AS-4 and AS-5 (Fig 1). This tilt is based on satellite altimetry derived SLA data produced by Segment Sol Multimission Altimetry and Orbitography (SSALTO) / Developing Use of Altimetry for Climate Studies (DUACS) and distributed by Archiving, Validation and Interpretation of Satellite Oceanographic Data (AVISO), with support from Centre National d'Etudes Spatiales

(CNES) (http://www.aviso.altimetry.fr/duacs/) and is referred to as 'DUACS SLA' data in this paper. In a way, tilting can be interpreted as a form of data assimilation in which observed SLA data are nudged into numerical model directly through open boundary. It is deemed insightful to qualitatively and quantitatively describe different possible contributions of residual to understand general physical and mathematical effects on the present barotropic numerical model SCSMC [1, 2]. Therefore to explicitly describe the contribution of these effects to residual, input variables for present GP implementation will include computed results of tidally and meteorologically driven numerical model that represents local hydrodynamics, and DUACS SLA that represents non-tidal water level contribution originating beyond the numerical model domain.

2.3. Implementation of GP

Table 1 summarizes inputs and parameters applied for inducing a model to describe the residual representation in Malacca Strait. Following this nomenclature the residual ($WL_{residual,s,t}$) of a particular station s at time t is defined by

$$WL_{residual,s,t} = WL_{model,s,t} - WL_{observation,s,t} \quad (1)$$

where $WL_{model,s,t}$ is water level computed by numerical model and $WL_{observation,s,t}$ is observed water level obtained from UHSLC dataset.

It is noted that numerical model applied here is SCSMC driven by tidal and meteorological forcing, and without any tilt. Another input variable $AS_{j,t}$ ($AS_{1,t}, AS_{2,t}, AS_{3,t}, AS_{4,t}$ *and* $AS_{5,t}$) represents DUACS SLA data at Andaman Sea boundary point location $j = 1,2,3,4$ and 5 representing AS-1, AS-2, AS-3, AS-4 and AS-5 (Fig 1), respectively. Output of GP would take the form of

$$WL_{residual,s,t} = function(WL_{model,s,t}, AS_{j,t}) \quad (2)$$

Referring to Table 1, function set excluded arithmetic operators such as hyperbolic tangent ($tanh$) to avoid highly non-linear component interactions which enhance risk of local optimization of model parameters. Data of input variables are available at daily resolution at 00:00 hour GMT+8 for period between years 1993-2006. Dataset is divided into two sets; one is period of ten years (1993 to 2002) for training of the GP model, and other is period of four years (2003 to 2006) for validation of GP model. Table 2 presents statistical parameters of residuals at each station in the Malacca Strait.

Table 1. Description of input and parameters used in genetic programming for inducing GP-residual model

Input or parameter	Value
Terminal set	$\{WL_{model,s,t}, AS_{j,t}\}$
Function set	$\{+, \times, -, \div, \exp, \text{power}, \log, \text{abs}\}$
Population model	Panmictic, generational, elitist
Selection method	Tournament selection
Population size	100
Initialization method	Grow method on size
Initial size of formulae	15
Crossover rate	40%
Mutation rate	5%
Maximum size of formula	45
Fitness measure	Root mean squared error
Stopping criterion	2 minutes of process time on 3.4GHz computer

Table 2. Statistical parameters of daily water level residual at each station in the Malacca Strait over 14 years (1993 to 2006)

Station	Minimum (m)	Maximum (m)	Mean (m)	Standard deviation (m)
Ko Taphao Noi	-0.612	0.394	-0.140	0.141
Langkawi	-0.774	0.225	-0.183	0.166
Penang	-0.603	0.380	-0.124	0.163
Lumut	-0.610	0.382	-0.078	0.174
Kelang	-0.781	0.491	-0.109	0.215
Keling	-0.674	0.542	0.014	0.187
Kukup	-0.769	0.530	-0.092	0.238

3. Results and discussions

Table 3 summarizes output models of GP to represent residuals at each station in the Malacca Strait. Performance of each of these models during the validation period is presented in root mean squared error (RMSE), correlation coefficient and percentage improvement (%IMP) in terms of RMSE of total water level prediction in Table 4. Fig 2 and Fig 3 show associated scatter plots of observed water level versus water level of numerical model with GP-residual model correction.

Table 3. GP-residual model induced for each station in Malacca Strait

Station	GP-residual model
Ko Taphao Noi	$WL_{residual,KoTaphaoNoi,t} = -0.1475 - AS_{1,t}$
Langkawi	$WL_{residual,Langkawi,t} = -0.086252 - 0.966 * AS_{5,t} - 0.286376 * WL_{model,Langkawi,t}$
Penang	$WL_{residual,Penang,t} = -0.113975 - 0.754523 * AS_{5,t} + 0.226372 * WL_{model,Penang,t}$
Lumut	$WL_{residual,Lumut,t} = -0.072 - AS_{5,t}$
Kelang	$WL_{residual,Kelang,t} = -AS_{5,t}$
Keling	$WL_{residual,Keling,t} = -0.408908 * AS_{5,t} + 0.113708 * WL_{model,Keling,t} + 0.4 * AS_{5,t} * WL_{model,Keling,t}$
Kukup	$WL_{residual,Kukup,t} = -0.143727 - 0.311931 * AS_{4,t} - AS_{5,t} + 0.167161 * WL_{model,Kukup,t} + 0.557202 * AS_{4,t} * WL_{model,Kukup,t} + 0.557202 * AS_{4,t}^{2}$

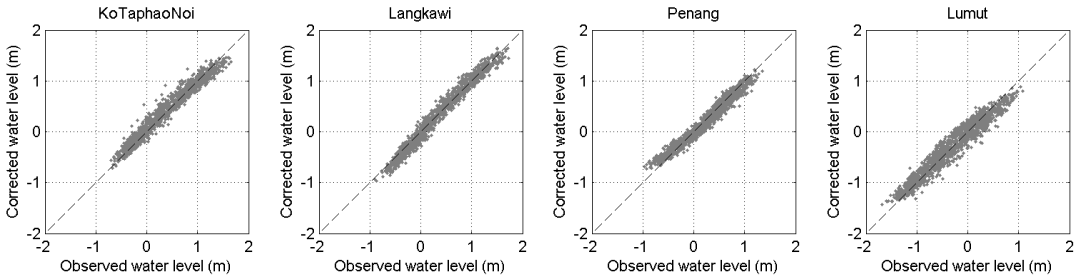

Fig 2. Scatter plots of observed water level and water level of numerical model with GP-residual model correction at (a) Ko Taphao Noi, (b) Langkawi, (c) Penang and (d) Lumut during the validation period

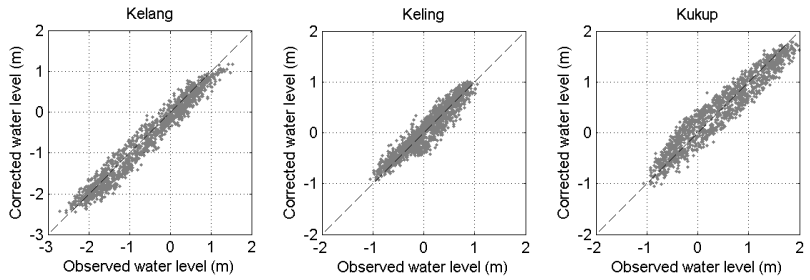

Fig 3. Scatter plots of observed water level and water level of numerical model with GP-residual model correction at (a) Kelang, (b) Keling and (c) Kukup during the validation period

Table 4. RMSE and correlation coefficient of total water level prediction with and without GP-residual model correction over the validation period

Station	Without GP-residual model correction		With GP-residual model correction		
	RMSE	Correlation coefficient	RMSE	Correlation coefficient	%IMP (RMSE)
Ko Taphao Noi	0.240	0.939	0.178	0.947	25.6
Langkawi	0.262	0.956	0.148	0.973	43.5
Penang	0.228	0.950	0.130	0.965	43.1
Lumut	0.257	0.912	0.214	0.930	16.8
Kelang	0.440	0.897	0.419	0.907	5.0
Keling	0.242	0.890	0.216	0.898	10.7
Kukup	0.394	0.891	0.332	0.901	15.7

Based on Table 3, it is noticeable that GP induced linear models to represent residual at the four stations located closest to the Andaman Sea boundary; Ko Taphao Noi to Lumut. Associated percentage improvement ranges between 16 to 44 percent. These linear GP-residual models show that the residual at Ko Taphao Noi and Lumut depends on DUACS SLA plus a constant, while Langkawi and Penang depend on same components plus an additional variable, in this case model water level. GP-residual model of Langkawi and Penang can be generalized to

$$WL_{residual,s,t} = \alpha + \lambda * AS_{j,t} + \gamma * WL_{model,s,t} \quad (3)$$

in which constant α can be interpreted as an error in local water level datum (or mean sea level), λ as contribution factor of SLA originating beyond the model domain, and γ as contribution factor of the model error due to missing forcing such as baroclinic and river discharge effect, and approximation error such as grid and bathymetry schematization. For stations Ko Taphao Noi and Lumut, GP-residual model can also be generalized to equation 3, but with $\gamma = 0$.

Referring to Table 4 GP-residual models for Langkawi and Penang provide the highest improvement in total water level prediction of more than 40 percent. Whereas GP-residual models for Ko Taphao Noi and Lumut show relatively lower improvement (less than 25 percent) in total water level prediction compared to Langkawi and Penang. Assuming residual in these four stations (Ko Taphao Noi to Lumut) takes form of equation 3, optimization of parameters α, λ and γ can be carried out to refine GP-residual models. Optimization problem is dealt better using another evolutionary algorithm technique; genetic algorithm (GA) [10, 11]. Optimizing parameters α, λ and γ of GP-residual models at the four stations is attempted using GA tool in MATLAB Global Optimization Toolbox over period between year 1993 to 2006. Table 5 shows GA-optimized parameters α, λ and γ of the GP-residual models. Values of GA optimized parameters do not deviate greatly (order of 0.001) from the original values induced by GP. Table 6 shows RMSE and correlation coefficient of total water level prediction based on GP-residual model correction with and without optimization of parameters using GA over period 1993 to 2006. RMSE improved

minimally (less than 0.009 m) at all four stations with GA optimized parameters. Little or no change in correlation coefficients is observed.

Another linear GP-residual model is the one for Kelang which solely depends on $AS_{5,t}$ without any coefficient (Table 3) with very small improvement of 5 percent which is also the lowest among all (Table 4). It is noted that tidal gauge of station Kelang is located at the mouth of river Sungei Puloh and could be sensitive to local river discharge. Without this discharge information, the residual could not be easily described. Regardless GP-residual model does imply that SLA originating in Andaman Sea may play limited role (about 5 percent) in the water level at Kelang.

Unlike dimensionally sound linear GP-residual models described above, the non-linear GP-residual models at Keling and Kukup present model which consists of multiplication of variables such as $AS_{j,t} * WL_{model,s,t}$ and ${AS_{j,t}}^2$ (Table 3), which resulted in improvements of only 10.7 and 15.7 percent, respectively (Table 4). It is noted that Keling and Kukup are stations located in complex tidal mixing zone with non-tidal hydrodynamics recognized as seasonally influenced by South China Sea and Andaman Sea [3, 19]. Therefore using $AS_{j,t}$ and $WL_{model,s,t}$ terms only may not be sufficient to fully describe residual. Furthermore, attempt to interpret every single term in these GP-residual models offers little physical meaning as they could be just a curve fitting solution. It is noted that it is not always possible to obtain physical meaning solely from data-driven techniques like GP. This is due to the fact that underlying errors attributed from assumptions and approximation made in numerical model will never cease to exist. If somehow these errors are known, numerical model can then be updated directly to improve prediction.

Table 5. GA optimized parameters in GP-residual models

Station	GA optimized parameters		
	α	λ	γ
Ko Taphao Noi	-0.1034	-0.7428	-0.0548
Langkawi	-0.1246	-0.9349	-0.2027
Penang	-0.1135	-0.8292	0.1767
Lumut	-0.0743	-1.0451	-0.0357

Table 6. RMSE and correlation coefficient of total water level prediction based on GP-residual model correction with and without optimization of parameters using GA over period 1993 to 2006

Station	Without optimization of parameters using GA		With optimization of parameters using GA	
	RMSE	Correlation coefficient	RMSE	Correlation coefficient
Ko Taphao Noi	0.161	0.962	0.158	0.962
Langkawi	0.121	0.983	0.112	0.983
Penang	0.136	0.966	0.132	0.966
Lumut	0.197	0.943	0.196	0.943

4. Conclusion

This paper described application of genetic programming (GP) for residual representation. GP is applied as a symbolic regression to determine a GP-residual model to describe local residual at stations in Malacca Strait using local computed water level of numerical model and DUACS tilt at Andaman Sea boundary points. Linear GP-residual models were induced at first five stations located in the northern Malacca Strait and are easily interpretable as different contributing components. Furthermore, Langkawi and Penang show highest improvement (more than 40 percent) in water level prediction corrected based on their corresponding GP-residual model, while other stations have improvement ranges between 5 to 25 percent. It is noted that GP exercise carried out for residual representation may not be inducing a global minimum search of the induced models. Induction of more accurate model could be possible.

Acknowledgement

The authors gratefully acknowledge the support and contributions of the Singapore–Delft Water Alliance (SDWA). The research presented in this work was carried out as part of the SDWA's "Must-Have Box" research program (R-264-001-003-272).

References

[1] S.H.X. Tay, A. Kurniawan, S.K. Ooi, V. Babovic, Further improvement of tidal representation in the South China Sea and the Southeast Asian waters. 2013 IAHR Congress. Chengdu. 2013.
[2] S.H.X. Tay, A. Kurniawan, S.K. Ooi, V. Babovic, Modelling Sea Level Anomalies in Malacca Strait. 36th IAHR Congress. The Hague. 2015.
[3] S.H.X. Tay, A. Kurniawan, S.K. Ooi, V. Babovic, Sea level anomalies in straits of Malacca and Singapore. Applied Ocean Research. 2016;58:104-17.
[4] Y. Sun, S.H.X. Tay, P. Sisomphon, P. Zemskyy, S.K. Ooi, H. Gerritsen, Study on the correlations between current anomalies and sea level anomaly gradients in Singapore region. 4th International Perspective on Water Resources & the Environment. Singapore. 2011.
[5] X. Wang, V. Babovic, Enhancing water level prediction through model residual correction based on Chaos theory and Kriging. International Journal for Numerical Methods in Fluids. 2014;75:42-62.
[6] A. Kurniawan, S.K. Ooi, V. Babovic, Improved sea level anomaly prediction through combination of data relationship analysis and genetic programming in Singapore Regional Waters. Computers & Geosciences. 2014;72:94-104.
[7] J.R. Koza, Genetic programming: on the programming of computers by means of natural selection: MIT press; 1992.
[8] V. Babovic, M. Keijzer, Genetic programming as a model induction engine. Journal of Hydroinformatics. 2000;2:35-60.
[9] V. Babovic, Introducing knowledge into learning based on genetic programming. Journal of Hydroinformatics. 2009;11:181-93.
[10] J.H. Holland, Genetic algorithms. Scientific american. 1992;267:66-72.
[11] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning: Addison-Wesley Longman Publishing Co., Inc.; 1989.
[12] H.-P.P. Schwefel, Evolution and optimum seeking: the sixth generation: John Wiley & Sons, Inc.; 1993.
[13] H.-G. Beyer, H.-P. Schwefel, Evolution strategies–A comprehensive introduction. Natural computing. 2002;1:3-52.
[14] A.E. Eiben, J.E. Smith, Introduction to evolutionary computing: springer; 2003.
[15] W. Banzhaf, P. Nordin, R.E. Keller, F.D. Francone, Genetic programming: an introduction: Morgan Kaufmann San Francisco; 1998.
[16] M. Keijzer, V. Babovic, Dimensionally aware genetic programming. In: Banzhaf W, Daida J, Eiben AE, Garzon MH, Honavar V, Jakiela M, Smith RE, editors. Gecco-99: Proceedings of the Genetic and Evolutionary Computation Conference. 1999. p. 1069-76.
[17] V. Babovic, Data Mining and Knowledge Discovery in Sediment Transport. Computer-Aided Civil and Infrastructure Engineering. 2000;15:383-9.
[18] M.J. Baptist, V. Babovic, J.R. Uthurburu, M. Keijzer, R.E. Uittenbogaard, A. Mynett, A. Verwey, On inducing equations for vegetation resistance. Journal of Hydraulic Research. 2007;45:435-50.
[19] A. Kurniawan, S.H.X. Tay, S.K. Ooi, V. Babovic, H. Gerritsen, Analyzing the physics of non-tidal barotropic sea level anomaly events using multi-scale numerical modelling in Singapore regional waters. Journal of Hydro-Environment Research. 2015;9:404-19.